第三方惩罚中的规范错觉：基于公正世界信念的解释*

杨莎莎[1]　陈思静[2]

([1] 上海大学经济学院，上海 200444) ([2] 浙江科技学院经济与管理学院，杭州 310023)

摘要 惩罚规范在一定程度上会影响个体的惩罚行为，但个体对惩罚规范的感知与实际规范之间可能存在差异，这被称为规范错觉。为了更好地从这一角度理解第三方惩罚，我们需要回答的是：第三方惩罚中是否存在规范错觉？如果存在，其方向如何？会对个体自身的惩罚行为产生何种影响？实验 1（$N = 449$）和实验 2（$N = 134$）的结果表明，在违规情境中，人们往往低估了他人的惩罚水平，这导致自身较低的惩罚行为。实验 3（$N = 164$）和实验 4（$N = 284$）进一步发现，较弱的公正世界信念导致人们对他人惩罚水平的低估，从而影响了自身的惩罚行为，而社会距离调节了公正世界信念对规范错觉的影响。上述结果表明，规范错觉会受到内部（公正世界信念）和外部（社会距离）两个参照点的影响，同时也在一定程度上说明第三方惩罚是一种注重维护规范的积极行为、而非注重个人收益的策略行为。

关键词 第三方惩罚，规范错觉，公正世界信念，社会距离

分类号 B849: C91

1 引言

非亲缘个体间的合作在保障人类社会顺利运行的过程中发挥着重要作用，为此我们发展出了合作的社会规范（social norm）（de Kwaadsteniet et al., 2007; Fehr & Schurtenberger, 2018）。但合作意味着个体需支付较高的成本（Rand, 2016），因此合作往往会随着自然选择（Gächter et al., 2017）和社会学习（Burton-Chellew et al., 2017）而崩溃。Fehr 和 Gächter（2002）提出的第三方惩罚（third-party punishment）——由利益无关者对违规者实施的惩罚——在一定程度上解释了人类社会中的合作行为：存在惩罚机制的情况下，1）个体预期违规行为会遭到严厉的惩罚，这降低了个体违规的动机（Molho et al., 2020）；2）个体认为他人也会因此减少违规，这又进一步提升了个体的合作行为（Lergetporer et al., 2014）；3）惩罚激活了受罚者内化的合作规范，从而提高了其在后续互动中的合作水平（陈思静, 邢懿琳等, 2021）。

收稿日期：2021-8-3

* 国家自然科学基金项目(71701185), 浙江省软科学项目(2020C35020)资助。

通信作者：陈思静, E-mail: chensijing@zust.edu.cn

来自现实社会的研究同样支持上述观点：不论是在公共资源管理中（Kosfeld & Rustagi, 2015），还是在战争与冲突中（Mathew & Boyd, 2011），惩罚都在很大程度上提高了团队的合作水平。

进一步需要回答的问题是什么因素导致人们做出惩罚行为？现有研究发现人们自身的惩罚行为往往依赖于他人的惩罚水平（FeldmanHall et al., 2018; Henrich & Boyd, 2001; Son et al., 2019），条件惩罚理论（conditional punishment theory）认为，只有在他人实施惩罚时个体才会实施惩罚（Huang et al., 2018; Molleman et al., 2019）；Li 等（2021）进一步指出，个体会根据他人在类似场合下的惩罚水平来实施自己的惩罚，他们将其称为惩罚的社会规范。上述观点的前提条件是个体能够充分获取他人的惩罚信息，但 Kiyonari 和 Barclay（2008）指出，在现实生活中这一点往往难以实现，因此我们需要进一步思考：无法充分获取他人惩罚信息时，惩罚行为究竟受到何种因素的影响？

Kamei（2020）指出了感知惩罚规范（即对他人惩罚行为的估计）的重要性，他们的研究表明，个体的惩罚行为与感知到的惩罚强度呈正相关。但个体对规范的感知与实际的规范之间可能存在差异（Rimal & Lapinski, 2015），这被称为规范错觉（normative misperception）（Cox et al., 2019）。出于对积极自我形象的追求（Leary, 2007），个体往往存在着自我服务偏差（self-serving bias）（Zhang et al., 2018），对他人的贬损恰好可以满足这一心理需求（Rau et al., 2019）。这也使得规范错觉总是表现为一种系统性的认知偏差：人们倾向于低估他人的积极行为，例如公众低估了他人的亲环境行为（Bouman et al., 2020; Bouman & Steg, 2019），青少年低估了同龄人对校园欺凌行为的反对程度（Dillon & Lochman, 2019），沙特阿拉伯的已婚男性低估了其他男性对女性劳动参与的支持（Bursztyn et al., 2020）；而倾向于高估他人的消极行为，如食物浪费（陈思静, 濮雪丽等, 2021）和酗酒（Amialchuk et al., 2019; Dumas et al., 2019）。因此规范错觉的存在减少了人们从事积极行为的可能性（Dempsey et al., 2018）。由于积极行为（如亲环境行为）以提高群体的福利水平为目的（Sawitri et al., 2015），但成本需要由个体承担（Keizer & Schultz, 2018），而第三方惩罚也具备上述特征（Enge et al., 2017），因此我们可以将其视为一种积极行为，并推测：个体倾向于低估惩罚的社会规范，并相应地减少了自身的惩罚行为。

然而针对合作的研究提出了相反的观点（Bicchieri et al., 2020; Goeschl et al., 2018），不少研究者发现人们倾向于高估公共物品博弈（public good game）中其他成员的贡献水平（Fischbacher & Gächter, 2010; Neugebauer et al., 2009），即存在高估的规范错觉。这一现象可以用 Fehr 和 Fischbacher（2004）提出的条件合作（conditional cooperation）来解释：受利己主义影响，大多数的条件合作者倾向于将自己的合作水平限制在他人平均水平之下，因为

这有利于将自身利益最大化。这种策略是否也存在于惩罚中？作为二阶合作行为（Ozono et al., 2016），惩罚的成本由个体承担收益却由群体共享，因此也被视为一种公共物品（Sasaki et al., 2015）。从这个角度来说，我们有理由认为惩罚中也可能存在类似的策略性动机，即高估他人惩罚行为，但这显然与第一种推论产生了矛盾。遗憾的是，目前尚无研究从规范错觉的视角来考察第三方惩罚，本文拟从该角度来探讨惩罚中规范错觉的方向、成因及其对个体惩罚行为的影响，从而为现有文献提供有益补充。

我们首先探讨的问题是：第三方惩罚中是否存在规范错觉？如果存在的话，其方向如何？上文的分析表明，规范错觉相关的多数研究认为个体倾向于高估他人的消极行为（陈思静，濮雪丽等, 2021; Amialchuk et al., 2019; Davis et al., 2019），而低估他人的积极行为（Anthenien et al., 2018; Dempsey et al., 2018），那么可以推测惩罚中存在着低估的规范错觉。如果参考来自合作领域的研究，更多地考虑策略性动机的影响，则会得到完全相反的推论。在此基础上，本文提出以下竞争性假设，并在实验 1 中进行检验：

H1a：人们系统性地低估了他人的第三方惩罚（低估的规范错觉）。

H1b：人们系统性地高估了他人的第三方惩罚（高估的规范错觉）。

需要进一步思考的是，规范错觉将如何影响惩罚行为？Grimm 等（2017）指出，对他人行为的感知往往会影响自身的行为，这可能是因为人们会通过声称自己的行为符合规范来证明其道德合理性（Schlag et al., 2015），例如对他人亲社会行为的低估会减少自身相应的行为（Ganz et al., 2020），而对他人物质使用行为（包括酗酒、吸烟和吸毒）的高估则增加了从事这类行为的可能性（Amialchuk et al., 2019）。我们推测，规范错觉对惩罚行为也有类似的影响。基于上述推理，本文提出以下竞争性假设，并在实验 1 和实验 2 中进行检验：

H2a：对他人惩罚行为的低估进一步降低了自身的惩罚行为。

H2b：对他人惩罚行为的高估进一步提升了自身的惩罚行为。

需要说明的是，在典型的第三方惩罚中，受罚者往往是因为违反了社会规范而受到他人惩罚，而违反社会规范的程度有高有低（Balafoutas et al., 2016），例如在独裁者博弈中，分配方案的公平程度在一定程度上体现了分配者是否违反社会规范以及违规的程度（Csukly et al., 2011），因此在检验上述两组竞争性假设时，我们还考察了违规程度是否影响人们对他人惩罚行为的错觉，作为对本研究结果的一个重要补充。

进一步的问题是：什么因素造成了这种偏差？正如胡金生等（2012）指出，公正判断更多地依赖于道德直觉而非外部社会线索，这也就意味着，当无法获取外部信息时，规范错觉可能依赖于个体的某种直觉或信念。Lerner（1965）提出的公正世界信念（belief in a just world,

BJW）理论提供了一种可能的解释。公正世界信念是一种比较稳定的个人特质（吴佩君, 李晔, 2014），它影响着个体遵守社会规范的意愿（姬旺华等, 2014; Lerner & Miller, 1978），这种信念越强，个体对违规行为的感知惩罚（Bai et al., 2014）与惩罚态度（Bègue & Bastounis, 2003）就越强，也就越愿意付出一定代价来惩罚违规者（Strelan et al., 2017）。换言之，公正世界信念影响了人们对惩罚的规范感知，而存在于集体中的真实规范在某个时点上是相对稳定的（Laninga-Wijnen et al., 2018），因此公正世界信念对惩罚规范感知的影响也可以视为对规范错觉（规范感知与真实规范之间的差异，具体定义参见 2.2 部分）的影响。基于上述推理，本文提出以下假设，并在实验 3 和实验 4 中进行检验：

H3：公正世界信念影响了第三方惩罚中的规范错觉。

此外，有研究者发现社会距离（social distance）——个体间的相似性或亲近度（徐杰等, 2017）——会影响个体对他人行为的关注（Charness & Gneezy, 2008），进而影响规范错觉（Cox et al., 2019; Kenney et al., 2017）。这也说明在不同的社会距离下，个体对同一事件的判断往往存在着偏差（徐杰等, 2017）。解释水平理论（Construal level theory, CLT）认为，在社会距离较远时，个体倾向于使用抽象的、普遍的观念去判断他人行为（Trope & Liberman, 2010），此时作为普遍观念的公正世界信念（Lerner & Miller, 1978）能对规范错觉产生更强的影响。但由于亲近他人（如亲友）的良好表现会给个体带来共同荣誉感（吴静珊, 王娜, 2017），随着社会距离的拉近，个体更有可能对他人的行为做出较高的道德评价（Tumasjan et al., 2011），即存在美化他人行为的功利主义动机（李明晖, 饶俪琳, 2017），此时社会距离的拉近可能会削弱公正世界信念对规范错觉的影响。基于上述推理，我们提出本文最后一个假设，并在实验 3 中进行检验：

H4：社会距离调节了公正世界信念对规范错觉的影响。

概括而言，本文包括以下 4 个实验：实验 1 主要探究第三方惩罚中的规范错觉及其对惩罚行为的影响；实验 2 则为实验 1 的稳健性检验，我们通过操纵个体对规范的感知，有针对性地改变了规范错觉，进而考察规范错觉与惩罚行为之间的因果关系；实验 3 探讨了公正世界信念与社会距离对规范错觉的影响；实验 4 进一步通过操纵公正世界信念来检验其与规范错觉之间的因果关系。

2 实验 1

2.1 被试

使用软件 G*Power3.1 进行的功效分析（power analysis）显示：取中等效应量 $d = 0.50$，

显著性水平 α = 0.05，配对样本 t 检验至少需要 54 名被试才能达到 95%（1 − β）的统计检验力；独立样本 t 检验每组至少需要 105 名被试才能达到 95%（1 − β）的统计检验力。由于实验 1 共包含两组配对样本、两组独立样本，按照样本要求量较大的独立样本 t 检验进行计算，至少需要 420 名被试。而实际参与实验 1 的被试为 449 名大学生，其中本科生占 77.06%，硕士研究生占 22.94%。被试平均年龄为 21.55 ± 1.56 岁，女性占 57.20%。被试的专业分布如下：理工类占 41.43%，人文类占 18.71%，社科类占 30.51%，艺术类及其他占 9.35%。实验开始前，实验者通过练习保证被试完全了解实验规则，并获得了所有被试的知情同意书（下同，不再赘述）。

2.2 设计

实验 1 为 4 组间因子设计（先惩罚后估计、先估计后惩罚、只需惩罚和只需估计），这一设计的目的是便于控制惩罚和估计的顺序对被试决策的干扰。实验 1 中惩罚和估计的测量均采用策略方法（strategy method）（Jordan et al., 2016; Volk et al., 2019），即所有被试都需要对 6 种不同的分配方案做出惩罚决策，并报告对其他成员平均惩罚水平的估计。采用这种方法的好处是，它完整地揭示了个体对不同分配方案的反应（Jordan et al., 2016）。实验 1 主要关注被试的规范错觉与惩罚行为。根据 Duong 和 Parker（2018）与 Sandstrom 等（2013）的建议，将某种分配方案下所有被试惩罚水平 PS 的平均值（M_{PS}）看作是存在于群体中的社会规范，而被试对他人惩罚水平的估计 PO 代表了被试在个人层面上对社会规范的感知，两者的差值即为惩罚的规范错觉。简言之，规范错觉的操作定义如下：规范错觉 ＝PO − M_{PS}。惩罚行为的操作定义则为被试针对不同分配方案实际实施的惩罚水平 PS。

需要说明的是，本文采用规范错觉（PO − M_{PS}）而非感知规范（PO）作为关键变量，主要是出于以下两方面的考虑。第一，不论是从统计还是定义的角度来看，规范错觉和感知规范都能够较好地解释个体的惩罚行为。第二，规范错觉相较于感知规范而言，可以表达出更为丰富的内涵：1）规范错觉能够更好地展示个体对他人惩罚水平的估计是否与实际水平发生偏离，这也是本文的重点关注之一；2）在解释惩罚行为时，采用规范错觉可以更直观地展示当规范错觉存在或不存在时个体惩罚行为的变化。

2.3 实验程序

实验 1 的范式为带有第三方惩罚的独裁者博弈，通过 z-Tree 上机实验完成（Fischbacher, 2007）。被试被随机分入上述四种实验条件，在整个实验过程中，被试均位于单独隔间内，相互之间无法交流。被试首先了解独裁者博弈的实验规则：分配者和接受者共同分配 10 代币，由分配者提出分配方案，接受者无法拒绝；共存在 6 种不同的分配方案，从 10-0（即保

留 10 代币，分配给接受者 0 代币）到 5-5；所有被试均作为第三方，各拥有 5 代币，面对 6 种不同的分配方案，被试可惩罚其认为不公平的分配者，惩罚规则为，被试每支付 1 代币即可扣减分配者 2 代币；被试需报告其自身的惩罚水平（即为了惩罚分配者而支付的代币数），并估计在每种分配方案中其他被试的平均惩罚水平。博弈共 1 轮，6 种分配方案按随机顺序出现，实验报酬为出场费加上随机抽取的 1 轮中被试剩余的代币数量。

2.4 结果与讨论

总体上被试的平均惩罚为 2.31 ± 1.61，平均估计为 2.17 ± 1.53。基于性别的平均数差异检验表明，男性和女性的惩罚水平不论是在整体上（$t(2050) = 0.14, p = 0.887$），还是在不同方案中（$t(340) = −0.51$~$0.75, ps = 0.455$~$0.988$），均不存在显著差异。然而男性对他人惩罚的估计显著高于女性（$t(1972) = 2.16, p = 0.031, d = 0.10$, 95%CI = [0.01, 0.29]），这种差异主要体现在 9-1 分配（$t(327) = 2.42, p = 0.016, d = 0.27$, 95%CI = [0.05, 0.49]）和 7-3 分配中（$t(327) = 2.02, p = 0.045, d = 0.23$, 95%CI = [0.005, 0.39]）。不同教育程度和专业下惩罚（$t(340) = −1.44$~$0.75, ps = 0.152$~0.467; $F(3, 338) = 0.94$~$2.33, ps = 0.074$~0.422）和估计（$t(327) = −0.13$~$1.17, ps = 0.245$~0.897; $F(3, 325) = 0.54$~$2.28, ps = 0.079$~0.658）的差异不显著，且年龄与惩罚和估计之间的相关系数也不显著（$r = −0.08$~$0.04, ps = 0.142$~0.803）。

根据上述四种实验条件将被试划分为四组，即先惩罚后估计（记为组 1，$n = 108$，包含惩罚 1 和估计 1）、先估计后惩罚（记为组 2，$n = 114$，包含惩罚 2 和估计 2）、只惩罚（记为组 3，$n = 120$，包含惩罚 3）和只估计（记为组 4，$n = 107$，包含估计 4）。表 1 给出了四组被试在面对不同分配方案时惩罚和估计的情况。

表 1 四种实验条件下惩罚和估计的描述性统计（$M \pm SD$）

组别	先惩罚后估计		先估计后惩罚		只惩罚	只估计
方案	惩罚 1	估计 1	惩罚 2	估计 2	惩罚 3	估计 4
10-0	4.19 ± 0.98	3.94 ± 1.08	4.14 ± 1.05	3.90 ± 1.09	4.28 ± 0.97	3.85 ± 1.11
9-1	3.46 ± 1.01	3.22 ± 1.01	3.48 ± 0.95	3.27 ± 0.99	3.50 ± 0.94	3.13 ± 1.03
8-2	2.82 ± 0.94	2.63 ± 0.91	2.80 ± 0.85	2.64 ± 0.93	2.83 ± 0.77	2.69 ± 0.94
7-3	2.00 ± 0.91	1.97 ± 0.85	2.16 ± 0.77	2.07 ± 0.83	2.01 ± 0.79	2.00 ± 0.90
6-4	1.06 ± 0.99	1.16 ± 0.83	1.18 ± 0.81	1.17 ± 0.79	1.17 ± 0.85	1.05 ± 0.71
5-5	0.11 ± 0.44	0.16 ± 0.50	0.20 ± 0.50	0.19 ± 0.44	0.18 ± 0.48	0.17 ± 0.42

对估计 1 和惩罚 1、估计 2 和惩罚 2、估计 4 和惩罚 3 进行平均数差异检验。估计 1 和惩罚 1 的配对样本 t 检验结果显示，在 10-0（$t(107) = -4.14$, $p < 0.001$, $d = -0.40$, 95%CI = [−0.37, −0.13]）、9-1（$t(107) = -4.33$, $p < 0.001$, $d = -0.42$, 95%CI = [−0.35, −0.13]）和 8-2（$t(107) = -3.19$, $p = 0.002$, $d = -0.31$, 95%CI = [−0.32, −0.07]）这三个分配方案中，个体对他人平均惩罚的估计显著低于实际的惩罚水平，即存在低估的规范错觉。然而，当分配方案为 7-3、6-4 和 5-5 时，估计和惩罚并无显著差异（$t(107) = -0.48$~1.52, $ps = 0.132$~0.633, $BF_{01} = 3.10$~8.38）。由于零假设显著性检验（null hypothesis significance testing, NHST）无法直接给出是否支持零假设的证据，因此我们用 JASP0.14.1 计算了相应的贝叶斯因子，结果显示当前数据更加支持零假设（胡传鹏等, 2018）：在后三种方案中不存在规范错觉。对估计 2 和惩罚 2 的检验显示了相似的结果：在 10-0（$t(113) = -3.76$, $p < 0.001$, $d = -0.35$, 95%CI = [−0.38, −0.12]）、9-1（$t(113) = -3.06$, $p = 0.003$, $d = -0.29$, 95%CI = [−0.35, −0.07]）和 8-2（$t(113) = -2.19$, $p = 0.031$, $d = -0.20$, 95%CI = [−0.30, −0.01]）中，存在低估的规范错觉；而在 7-3、6-4 和 5-5 中，估计和惩罚并没有显著差异（$t(113) = -1.22$~-0.13, $ps = 0.227$~0.897, $BF_{01} = 4.70$~9.54），贝叶斯因子分析也说明当前数据更加支持零假设。类似地，估计 4 和惩罚 3 的独立样本 t 检验结果显示，当分配方案为 10-0（$t(225) = -3.13$, $p = 0.002$, $d = -0.41$, 95%CI = [−0.71, −0.16]）、9-1（$t(225) = -2.82$, $p = 0.005$, $d = -0.37$, 95%CI = [−0.63, −0.11]）和 8-2（$t(225) = -2.05$, $p = 0.042$, $d = -0.27$, 95%CI = [−0.46, −0.01]）时，被试显著低估了他人的惩罚行为；而在 7-3、6-4 和 5-5 中，估计和惩罚没有显著差异（$t(225) = -1.15$~-0.07, $ps = 0.253$~0.941, $BF_{01} = 3.71$~6.87），贝叶斯因子显示当前数据更加支持零假设。图 1、图 2 和图 3 直观地展示了不同分配方案中估计和惩罚的差异情况。

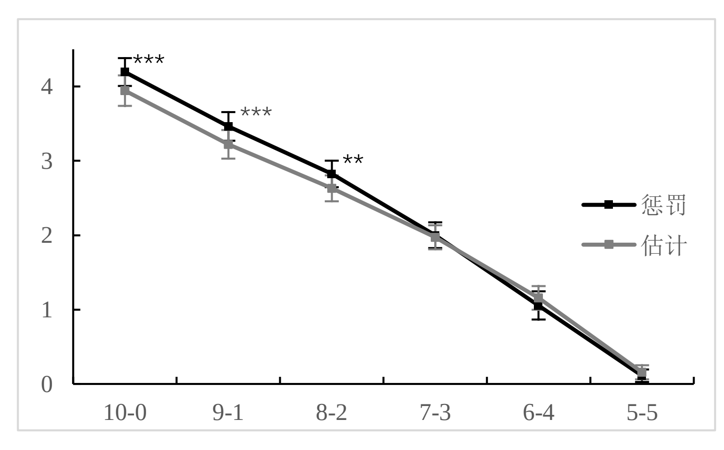


图 1 先惩罚后估计中估计 1 和惩罚 1 的比较

注：$n = 108$，$^{***}p < 0.001$，$^{**}p < 0.01$。

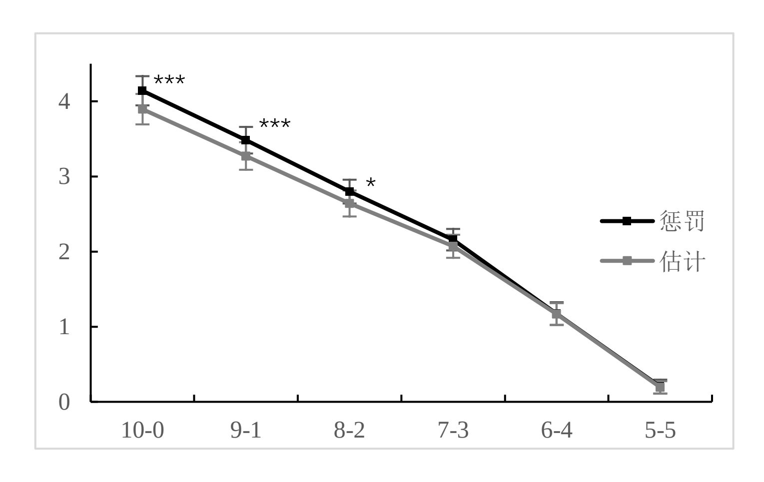


图 2 先估计后惩罚中估计 2 和惩罚 2 的比较

注：$n = 114$，$^{***}p < 0.001$，$^{*}p < 0.05$。

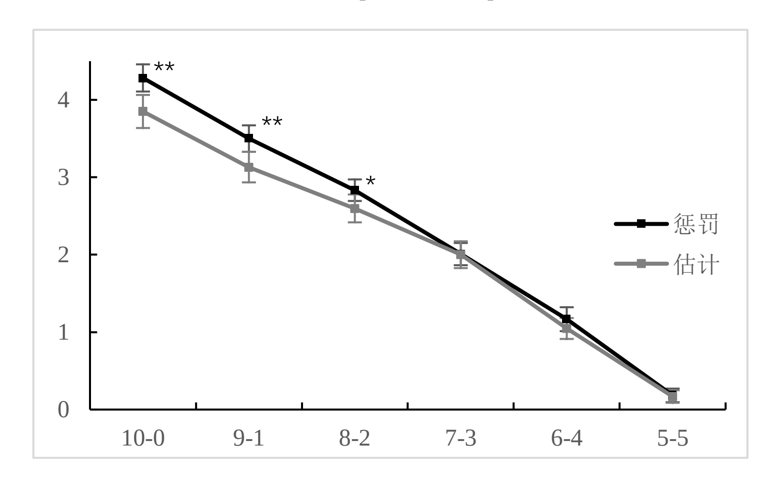


图 3 只估计和只惩罚中估计 4 和惩罚 3 的比较

注：$n_3 = 120$，$n_4 = 107$，$^{**}p < 0.01$，$^{*}p < 0.05$。

先前有相当文献表明在不同文化语境中人们对于什么样的分配方案算是违规/合作有高度稳定的看法，即分配给对方的金额约小于 30%是一种违规行为（Csukly et al., 2011; Fehr & Fischbacher, 2003），且有学者认为这种在划分标准上的稳定性具有一定的生物学基础（Wallace et al., 2007）。以初始金额（10 代币）30%计算，3 代币为分界点，也就是 10-0、9-1 和 8-2 的分配方案可被认为是不公平的。因此上述结果可总结为：1）面对不公平的分配方案时，被试往往低估了其他惩罚者的平均惩罚强度，即存在显著的规范错觉；2）当分配方案公平时，并不存在显著的规范错觉。这意味着第三方惩罚中确实存在低估的规范错觉，但这种错觉仅仅存在于分配方案不公平时。这部分地支持了本文的假设 1a。此外，由于对违规者的惩罚是一种积极的行为（Li et al., 2018），因此实验 1 的结果也在一定程度上说明，在惩罚行为中，个体也会产生与亲环境行为（Bouman & Steg, 2019）等相似的规范错觉，即低

估别人的积极努力。

为排除惩罚和估计的相对顺序对错觉可能产生的影响，使用单因素方差分析，比较了惩罚 1、惩罚 2 和惩罚 3，以及估计 1、估计 2 和估计 4。结果显示，被试的平均惩罚（$F(2, 339) = 0.04\sim1.32$, $ps = 0.268\sim0.959$, $BF_{01} = 9.30\sim30.11$）和平均估计（$F(2, 326) = 0.06\sim0.80$, $ps = 0.450\sim0.941$, $BF_{01} = 14.53\sim28.58$）在不同实验条件下均不存在显著差异，贝叶斯因子分析的结果也表明当前数据更加支持零假设。这说明惩罚和估计的测量顺序并不会影响被试的惩罚行为以及对他人惩罚行为的估计，惩罚中的规范错觉并非是由测量顺序或两种测量相互的影响而造成的。

为进一步探究规范错觉对被试自身惩罚行为的影响，进行规范错觉对惩罚行为的回归分析。根据前述分析，分配是否公平会影响规范错觉，因此我们在分析中控制了不公平程度，其定义是实际分配与均等分配的偏离程度，记 10-0 为 5（表示完全不公平），记 5-5 为 0（表示完全公平）。考虑到上述单因素方差分析的结果，即不同实验条件不会对惩罚和估计产生显著影响，因此在分别对组 1 和组 2 进行回归分析后，将两组数据合并，作为结果稳健性的检验。结果如表 2 所示。

表 2 规范错觉和不公平程度对惩罚行为的回归分析

变量	组 1（$n = 648$）				组 2（$n = 684$）				合并（$N = 1332$）			
	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*
不公平程度	0.81***	0.02	0.87***	0.01	0.78***	0.02	0.81***	0.01	0.80***	0.01	0.84***	0.01
规范错觉			0.79***	0.03			0.65***	0.03			0.72***	0.02
Adj-R^2	0.70***		0.88***		0.71***		0.84***		0.71***		0.86***	
ΔR^2			0.18***				0.13***				0.15***	

注：表中回归系数为标准回归系数，*** $p < 0.001$。

由表 2 可知，不公平程度对惩罚行为存在显著的正向影响，加入规范错觉后，R^2 分别增加了 0.18、0.13 和 0.15，说明规范错觉能显著地影响惩罚行为，在控制不公平程度的影响后，仍然可以解释因变量变异的 18%、13%和 15%。这也就意味着，对他人惩罚的低估将降低自身的惩罚行为，这为假设 2a 提供了初步证据。为了更直观地展示在违规情境中规范错觉对惩罚行为的影响，我们将组 1 和组 2 合并后，根据三种不公平的分配方案进行规范错觉

对惩罚行为的回归分析，结果如表 3 所示；图 4 则十分直观地展现了在不同分配方案下规范错觉（横轴）和自身惩罚行为（纵轴）之间的联系（气泡大小表示该位置上的被试数量）。

表 3 规范错觉对惩罚行为的回归分析

方案	*B*	*SE*	β	LLCI	ULCI	Adj-R^2
10-0	0.75***	0.04	0.80	0.68	0.82	0.64***
9-1	0.76***	0.04	0.78	0.68	0.84	0.60***
8-2	0.68***	0.05	0.70	0.58	0.77	0.48***

注：$n = 222$，$^{***}p < 0.001$。

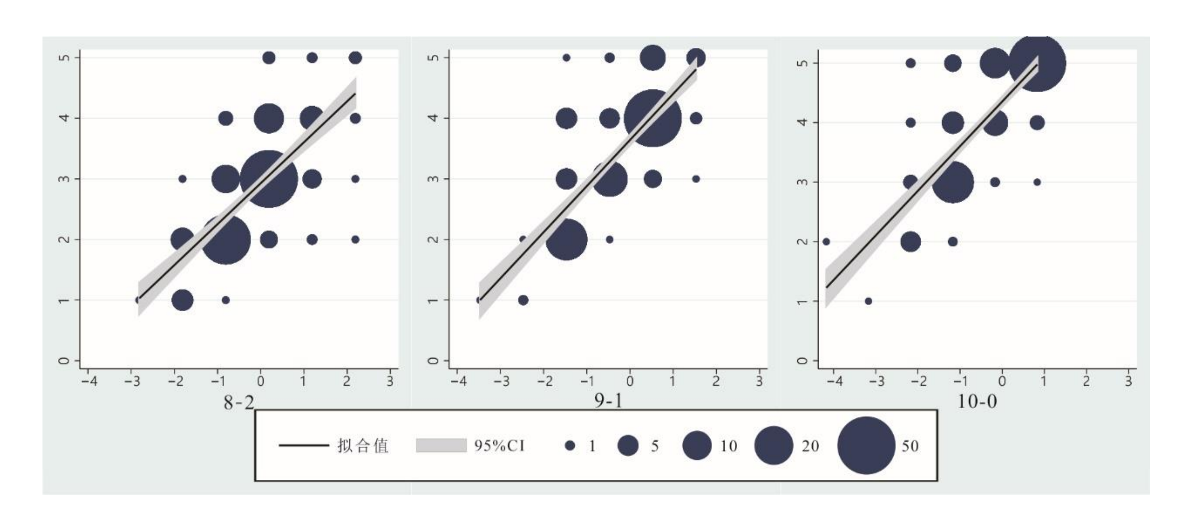


图 4 不同分配方案下规范错觉与惩罚行为的气泡图

作为补充说明，我们还通过不公平程度对规范错觉的回归分析探究了违规情境下不公平程度对规范错觉的影响（考虑到在公平分配方案中不存在显著的规范错觉，因此仅选取 10-0、9-1 和 8-2 这三个不公平分配方案），结果显示：违规情境下不公平程度对规范错觉的回归系数不显著（$\beta = -0.03, t = -0.76, p = 0.450$）。这也意味着，分配方案是否公平影响了被试的规范错觉，即只有在不公平分配方案中才存在规范错觉，但当分配方案为不公平时，不公平程度对规范错觉的影响却不显著。

不难注意到，虽然绝大多数被试的惩罚行为依赖于规范错觉，但仍有小部分被试在高估他人惩罚行为的情况下，实施了较低的惩罚（本文称其为二阶搭便车者）；此外，也有部分被试虽然低估了他人的惩罚行为，自身却表现出了较高的惩罚水平（本文称其为强互惠者）。

进一步探究强互惠者和二阶搭便车者在群体中所占比重可以加深对第三方惩罚中规范错觉的理解。我们以平均惩罚为分类依据，将被试划分为高惩罚-高错觉、高惩罚-低错觉、低惩罚-低错觉和低惩罚-高错觉四类，在合并组 1 和组 2 前三个分配方案的数据后，绘制了如图 5 所示的象限图。强互惠者（高惩罚-低错觉）落在第二象限，占总人数的 9.01%，二阶搭便车者（低惩罚-高错觉）落在第四象限，占总人数的 3.15%。强互惠者的人数远高于二阶搭便车者，这也意味着第三方惩罚具有较强的互惠性，即便是在单次匿名博弈中，二阶搭便车行为也不常见。上述结果似乎表明 Fehr 和 Fischbacher（2004）提出的条件合作策略并不能很好地推广至第三方惩罚领域，在不公平的分配方案中，我们也许更应该将第三方惩罚视为一种注重维护规范的积极行为、而非注重个人收益的策略行为。

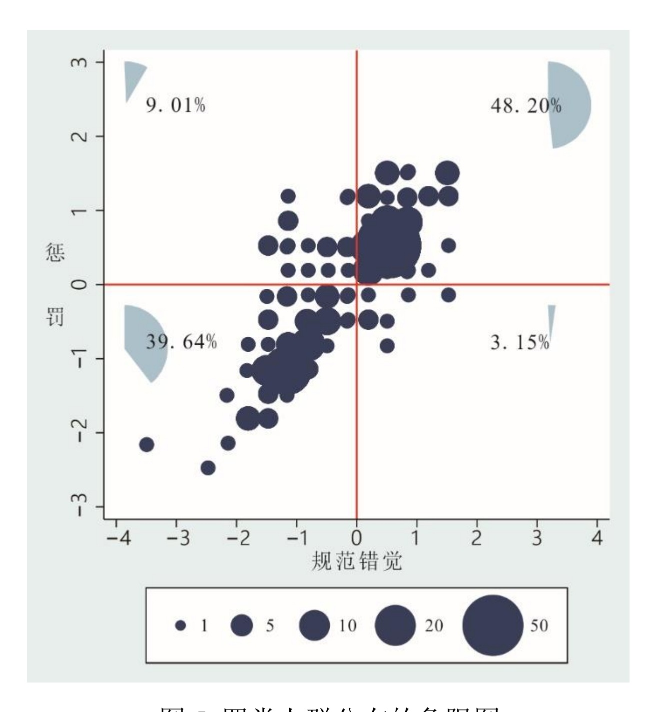


图 5 四类人群分布的象限图

3 实验 2

虽然实验 1 的结果从统计分析上说明规范错觉影响了被试的惩罚行为，但两者是否存在因果关系仍需更直接的检验。因此实验 2 主要通过有针对性的规范提示来改变被试的规范错觉，进一步考察两者间的因果关系，从而与实验 1 的结果形成印证。

3.1 被试

使用软件 G*Power 3.1 进行的功效分析显示：取中等效应量 f^2 = 0.15，显著性水平 α = 0.05，多元回归至少需要 119 名被试才能达到 95%（1 − β）的统计检验力；取中等效应量 d

= 0.50，显著性水平 α = 0.05，配对样本 t 检验至少需要 54 名被试才能达到 95%（1 − β）的统计检验力（实验 2 共包含两组配对样本，至少需要 108 名被试）。按照被试要求量较大的多元回归进行计算，至少需要 119 名被试，而实际参与实验 2 的被试为 134 名大学生，其中本科生占 72.39%，硕士研究生占 27.61%。被试平均年龄为 21.19 ± 1.82 岁，女性占 55.22%。被试的专业分布如下：理工类占 41.79%，人文类占 23.13%，社科类占 26.12%，艺术及其他占 8.96%。

3.2 变量与设计

实验 2 根据规范提示的方向设置了两种实验条件：提示高估组（n = 69）和提示低估组（n = 65），即告知被试高估/低估了他人的平均惩罚水平。自变量为规范提示，提示高估组记为 0，提示低估组记为 1。因变量和中介变量分别为进行规范提示后的惩罚水平（PS）和规范错觉（PO − M_{PS}），操作定义与实验 1 类似。但实验 2 中被试只需针对 8-2 的分配方案进行估计和实施惩罚，主要是出于以下两个方面的考虑：1）若此处选择重复博弈，前一轮博弈的规范提示可能会影响后一轮博弈中提示前的惩罚和估计；2）实验 1 的结果显示，对于三种不公平分配方案而言，分配方案的不公平程度在很大程度上不会影响规范错觉的强弱，因此本文参考现有文献选取了 8-2 的分配结果（Chen et al., 2020; Przepiorka & Liebe, 2016）。需要说明的是，在计算两组被试的规范错觉时，减去的平均惩罚（M_{PS}）为组内平均值而非总平均值。

3.3 实验程序

实验 2 的范式为带有第三方惩罚的独裁者博弈，通过 z-Tree 上机实验完成（Fischbacher, 2007）。被试被随机分入上述两种实验条件。实验开始后，被试首先了解独裁者博弈的实验规则：分配者和接受者共同分配 10 代币，由分配者提出分配方案，接受者无法拒绝；第三方拥有 5 个初始代币，可惩罚分配者，规则为第三方每支付 1 代币分配者的收益就降低 2 代币。随后向其展示第一轮博弈的分配结果（8-2 分配），被试需估计参加此次实验的其他第三方在面对该分配方案时的平均惩罚水平，并报告自己所选择的惩罚水平。提示高估组的被试被告知其高估了他人的平均惩罚，提示低估组的被试则被告知其低估了他人的平均惩罚。被试在收到上述规范提示后调整自己的估计和惩罚。博弈共 1 轮，实验报酬为出场费加上被试手中剩余的代币数量。

3.4 结果与讨论

不同性别（$t(132)$ = −0.46~0.77, ps = 0.443~0.765）、教育程度（$t(132)$ = −0.73~0.33, ps = 0.465~0.947）和专业类型（$F(3, 130)$ = 0.06~0.40, ps = 0.754~0.981）下，被试在规范提示前

后的惩罚、估计和规范错觉均无显著差异，且年龄和上述变量之间的相关系数也不显著（$r = −0.04$~0.03, $ps = 0.680$~0.914）。表 4 给出了两组被试在规范提示前后上述各变量的描述性统计。通过比较规范提示前估计和惩罚的差异，可以为实验 1 提供稳健性检验。结果显示：在规范提示之前，提示高估组（$t(68) = −2.56$, $p = 0.013$, $d = −0.31$, 95%CI = [−0.31, −0.04]）和提示低估组（$t(64) = −2.55$, $p = 0.013$, $d = −0.32$, 95%CI = [−0.33, −0.04]）中，均存在低估的规范错觉。

表 4 两组被试惩罚、估计和规范错觉的描述性统计（$M \pm SD$）

变量	提示高估组（$n = 69$）		提示低估组（$n = 65$）	
	规范提示前	规范提示后	规范提示前	规范提示后
惩罚	2.83 ± 0.77	2.23 ± 0.79	2.83 ± 0.74	3.28 ± 0.84
估计	2.65 ± 0.74	1.90 ± 0.86	2.65 ± 0.74	3.52 ± 0.81
规范错觉	−0.17 ± 0.74	−0.33 ± 0.86	−0.18 ± 0.74	0.25 ± 0.81

由于实验 2 的目的在于通过操纵估计来改变被试的规范错觉，进而影响其惩罚行为，即为了在规范错觉和惩罚行为之间建立因果关系，从而形成对实验 1 的补充说明，因此首先需要检验规范提示的操作有效性。需要说明的是，在规范提示前，两组被试的惩罚、估计和规范错觉均不存在显著差异（$t(132) = −0.04$~0.09, $ps = 0.932$~0.971, $BF_{01} = 5.39$~5.40），贝叶斯因子的结果也显示当前数据更加支持零假设，这也意味着本文较好地控制了两组被试在规范提示前的差异。而通过对两组被试在提示前后的规范错觉进行配对样本 t 检验，我们发现：在提示高估组中，提示后低估的规范错觉显著增强（$t(68) = −2.20$, $p = 0.032$, $d = −0.26$, 95%CI = [−0.30, −0.01]），提示低估组则呈现出相反的结果（$t(64) = 5.35$, $p < 0.001$, $d = 0.66$, 95%CI = [0.27, 0.59]）。这说明通过规范提示改变被试规范错觉的操作是有效的。进一步比较提示后的估计和惩罚可以更好地揭示规范错觉的变化情况，结果显示：在提示高估组中，对他人惩罚行为的估计仍然低于实际惩罚（$t(68) = −4.54$, $p < 0.001$, $d = −0.55$, 95%CI = [−0.48, −0.19]），存在低估的规范错觉；而在提示低估组中，对他人惩罚行为的估计显著高于实际惩罚（$t(64) = 3.11$, $p = 0.003$, $d = 0.39$, 95%CI = [0.09, 0.40]），存在高估的规范错觉。提示后两组的规范错觉存在显著差异（$t(132) = −4.01$, $p < 0.001$, $d = −0.70$, 95%CI = [−0.87, −0.29]）。上述结果综合表明，当被试得知自己高估了他人的平均惩罚后，对他人惩罚行为的低估程度会进一步

扩大；而当被试得知自己低估了他人的平均惩罚后，规范错觉从低估转变为高估。

其次，为探究规范错觉与惩罚行为间的因果关系，对规范提示前后的惩罚行为进行配对样本 t 检验，结果显示：在提示高估组中，规范提示后的惩罚行为显著低于提示前（$t(68) = -7.89$, $p < 0.001$, $d = -0.95$, 95%CI = [−0.74, −0.44]），提示低估组则恰好相反（$t(64) = 5.87$, $p < 0.001$, $d = 0.73$, 95%CI = [0.29, 0.60]）。独立样本 t 检验的结果也显示，提示高估组的惩罚行为显著低于提示低估组（$t(132) = -7.43$, $p < 0.001$, $d = -1.29$, 95%CI = [−1.32, −0.77]）。这在一定程度上说明，规范错觉确实影响了个体惩罚行为。具体来说，低估的规范错觉越弱，个体的惩罚水平越高。为增强结果的稳健性，需进一步检验规范提示-规范错觉-惩罚行为的中介模型。以规范提示为自变量，惩罚行为为因变量，规范错觉为中介变量，进行逐步回归，结果显示：规范提示能够直接影响被试的惩罚行为（$\beta = 0.54$, $p < 0.001$, 95%CI = [0.77, 1.32]）；而在加入规范错觉后，R^2 增加了 0.36，规范提示的回归系数下降但仍然显著（$\beta = 0.33$, $p < 0.001$, 95%CI = [0.44, 0.85]），且规范错觉的回归系数显著为正（$\beta = 0.64$, $p < 0.001$, 95%CI = [0.58, 0.81]），这说明规范错觉起到了部分中介作用。进一步使用 Preacher 和 Hayes（2004）所开发的 PROCESS3.3 检验上述中介模型（Bootstrap $N = 5000$, Model = 4），结果仍然显示：规范提示对惩罚行为的直接效应（effect = 0.64, $SE = 0.10$, LLCI = 0.44, ULCI = 0.85）和通过规范错觉影响惩罚行为的间接效应（effect = 0.40, $SE = 0.11$, LLCI = 0.20, ULCI = 0.61）均显著，且间接效应占总效应的 38.49%。综上所述，分析结果在一定程度上支持了规范错觉与惩罚行为之间的因果关系，从而形成对实验 1 结果的有益补充。

4 实验 3

实验 3 主要考察公正世界信念与规范错觉间的关系，以及感知社会距离在这一路径中的调节作用。

4.1 被试

使用软件 G*Power3.1 进行的功效分析显示：取中等效应量 $d = 0.50$，显著性水平 $\alpha = 0.05$，配对样本 t 检验至少需要 54 名被试才能达到 95%（$1 - \beta$）的统计检验力；取中等效应量 $f^2 = 0.15$，显著性水平 $\alpha = 0.05$，多元回归中需要 119 名被试才能达到 95% （$1 - \beta$）的统计检验力。按照样本要求量较大的多元回归进行计算，至少需要 119 名被试，而实际参与实验 3 的被试为 164 名大学生，其中本科生占 78.66%，硕士研究生占 21.34%。被试平均年龄为 21.19 ± 1.59 岁，女性占 53.05%。被试的专业分布如下：理工类占 37.80%，人文类占 16.46%，社科类占 35.98%，艺术类及其他占 9.76%。

4.2 变量与设计

实验 3 为被试内设计，自变量、因变量、中介变量和调节变量分别如下。

自变量 公正世界信念可以划分为两个维度：个人公正世界信念（personal belief in a just world）和一般公正世界信念（general belief in a just world）（Wu et al., 2011）。前者即相信世界对自身而言是公正的，往往预测了个体的主观幸福感（Sutton & Winnard, 2007）；后者则是相信世界对他人而言是公正的，也被称为他人公正世界信念（belief in a just world to others，本文将其统称为一般公正世界信念）（周春燕, 郭永玉, 2013），主要影响个体对社会现象的判断以及对社会环境的解释（Testé & Perrin, 2013）。由于研究者常将后者与感知惩罚（Bai et al., 2014）和惩罚态度（Bègue & Bastounis, 2003）相联系，因此将一般公正世界信念作为本文的自变量。我们参考了 Wu 等（2011）以及刘广增等（2020）的研究，通过 6 个题项来测量被试的一般公正世界信念（Cronbach's α = 0.85，下文简称为公正世界信念）。如“我认为这个世界基本上是公正的”和“我确信公正总是可以战胜不公正”等（详见附录 1），均为 6 点 Likert 量表：1 = 完全不同意；6 = 完全同意。得分越高说明公正世界信念越强。

因变量 实验 3 的因变量为惩罚水平，被试需针对按照随机顺序出现的不公平分配方案（10-0、9-1 和 8-2）做出惩罚，即在 0~5 中选择一个数字代表自己愿意用于惩罚的金额，记为 PS。

中介变量 实验 3 的中介变量为规范错觉，操作定义同实验 1。

调节变量 我们用改编自 Jones（2004）的 4 个题项（5 点计分，1 表示完全不同意，5 表示完全同意）测量了被试对于在此次实验中扮演第三方的大多数人的感知社会距离（Cronbach's α = 0.77，下文简称社会距离），如“我认为我和他们存在一些共同之处”和“在有些事情上，我会和他们有相同的看法”等（详见附录 2）。得分越高说明社会距离越近。

4.3 实验程序

实验 3 的范式为带有第三方惩罚的独裁者博弈，通过 z-Tree 上机实验完成（Fischbacher, 2007）。实验开始后，首先通过前述 6 个题项测量被试的公正世界信念，随后被试需了解博弈规则（同实验 2）。博弈共 3 轮，按照随机顺序出现 10-0、9-1 和 8-2 这三种不公平的分配方案，在每一轮博弈中，被试需估计参加此次实验的其他第三方的平均惩罚，并报告自己的惩罚决策。3 轮博弈结束后，我们通过前述 4 个题项测量了被试对于其他第三方的感知社会距离。实验报酬为出场费加上随机抽取的 1 轮中被试剩余的代币数量。

4.4 结果与讨论

不同性别（$t(162) = -1.17$~0.31, $ps = 0.244$~0.901）、教育程度（$t(162) = 0.31$~1.36, $ps =$

0.176~0.758）和专业类型（$F(3, 160) = 0.31$~1.05, $ps = 0.370$~0.815）下被试的惩罚和估计行为无显著差异，且年龄与惩罚和估计之间的相关系数也不显著（$r = -0.14$~-0.04, $ps = 0.082$~0.607）。各变量的描述统计与相关系数如表 5 所示。

表 5 变量描述性统计与相关系数

变量	M	SD	1	2	3	4	5	6	7
1 惩罚 10-0	4.21	1.04							
2 惩罚 9-1	3.56	1.00	0.87***						
3 惩罚 8-2	2.85	0.90	0.75***	0.85***					
4 规范错觉 10-0	−0.27	1.11	0.78***	0.70***	0.63***				
5 规范错觉 9-1	−0.26	1.01	0.72***	0.72***	0.62***	0.89***			
6 规范错觉 8-2	−0.20	0.94	0.67***	0.68***	0.67***	0.80***	0.86***		
7 公正世界信念	4.15	0.87	0.20**	0.19*	0.28***	0.29***	0.27***	0.31***	
8 社会距离	3.82	0.52	0.16*	0.14	0.12	0.26***	0.22**	0.24**	0.44***

注：$N = 164$，***$p < 0.001$，**$p < 0.01$，*$p < 0.05$。

4.4.1 中介效应检验

估计和惩罚的配对样本（$N = 164$）t 检验结果仍然显示，在 10-0（$t(163) = -4.88, p < 0.001, d = -0.38$, 95%CI = [−0.39, −0.16]）、9-1（$t(163) = -4.43, p < 0.001, d = -0.35$, 95%CI = [−0.38, −0.15]）和 8-2（$t(163) = -3.33, p = 0.001, d = -0.26$, 95%CI = [−0.31, −0.08]）这三个分配方案中，均存在低估的规范错觉，这与实验 1 的结果相吻合。此外，实验 1 的分析结果表明，在违规情境下，不公平程度对规范错觉无显著影响（$\beta = -0.03, t = -0.76, p = 0.450$）；我们在实验 3 中进行了类似的回归分析，得到了一致的结果（$\beta = -0.03, t = -0.70, p = 0.486$）。基于上述结果，我们在实验 3 中选择将三种方案进行合并分析，以被试在三种方案中的平均惩罚（3.54 ± 0.92）和平均规范错觉（-0.24 ± 0.97）作为惩罚和规范错觉的代理变量。为探究形成规范错觉的原因，我们以公正世界信念为自变量，规范错觉为因变量，进行了回归分析。结果显示，公正世界信念的回归系数显著为正（$\beta = 0.30, t = 4.04, p < 0.001$, 95%CI = [0.17, 0.50]），即公正世界信念越弱，低估的规范错觉就越强。假设 3 得到部分地验证。

结合实验 1 和实验 2 的结果——低估的规范错觉降低了被试的惩罚行为，进一步的问

题是公正世界信念是否能够通过影响规范错觉进而影响惩罚行为。我们以公正世界信念为自变量，惩罚行为为因变量，规范错觉为中介变量，进行逐步回归。表 6 的结果显示，公正世界信念对惩罚行为具有显著的正向影响，加入规范错觉后，R^2 增加了 0.54，且公正世界信念的回归系数变得不显著，这说明规范错觉在公正世界信念与惩罚行为之间起到了主要的中介作用（温忠麟, 叶宝娟, 2014）。进一步使用 Preacher 和 Hayes（2004）所开发的 PROCESS3.3 检验中介效应（Bootstrap $N = 5000$，Model = 4），结果显示：公正世界信念对惩罚行为的直接效应不显著（effect = 0.003, SE = 0.06, LLCI = −0.11, ULCI = 0.11），而通过规范错觉影响惩罚行为的间接效应均显著（effect = 0.25, SE = 0.07, LLCI = 0.12, ULCI = 0.39）。这与逐步回归的结果类似，即公正世界信念主要是通过个体的规范错觉来影响惩罚行为。

表 6 逐步回归对主效应和中介效应的检验（因变量：惩罚行为）

变量	M_1				M_2			
	β	*SE*	LLCI	ULCI	β	*SE*	LLCI	ULCI
自变量								
公正世界信念	0.24^{**}	0.08	0.09	0.41	0.003	0.06	−0.11	0.11
中介变量								
规范错觉					0.77^{***}	0.05	0.64	0.84
Adj-R^2	0.05^{**}				0.59^{***}			
ΔR^2					0.54^{***}			

注：$N = 164$，$^{***}p < 0.001$，$^{**}p < 0.01$。

4.4.2 调节效应检验

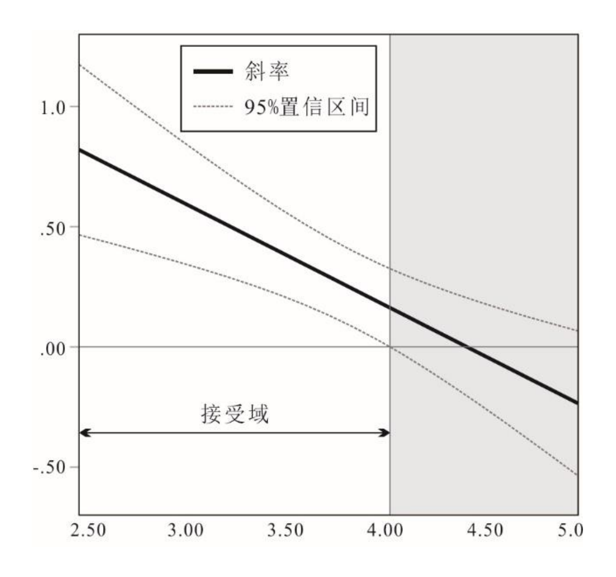
不少研究者指出，社会距离会影响个体对他人行为的关注（Charness & Gneezy, 2008）与评价（Tumasjan et al., 2011），进而影响规范错觉（Kenney et al., 2017; Cox et al., 2019）。因此进一步探究社会距离的调节效应有助于我们厘清惩罚中规范错觉形成的原因及其影响因素。以公正世界信念（BJW）为自变量，规范错觉为因变量，社会距离（SD）为调节变量，通过逐步回归来检验社会距离的调节效应，为降低共线性及方便比较，将各变量进行标准化处理。结果如表 7 所示：加入公正世界信念与社会距离的交互项（BJW×SD）后，R^2 增加了 0.07，且交互项的系数显著为负（β = −0.26, $p < 0.001$, 95%CI = [−0.33, −0.10]），说明社会距离在公正世界信念影响规范错觉的过程中起到了显著的负向调节作用。由于在本文中，题

项得分越高表明社会距离越近，因此上述结果可以描述为：在社会距离较远时，公正世界信念对规范错觉存在显著的影响；随着社会距离的不断拉近，这种影响就被削弱了。为更清晰地显示社会距离的调节作用，用 Johnson-Neyman 法进一步量化分析社会距离对公正世界信念与规范错觉关系的影响，并检验调节效应的统计显著区间，结果如图 6 所示。当社会距离得分小于 4 时，公正世界信念影响规范错觉的回归斜率置信区间为正且不包含 0，此时公正世界信念越强，低估的规范错觉就越弱；而当社会距离得分大于 4 时，置信区间包含 0 点，此时公正世界信念对规范错觉的影响并不显著。这较好地验证了假设 4。

表 7 逐步回归对调节效应的检验（因变量：规范错觉）

变量	M_1				M_2			
	β	*SE*	LLCI	ULCI	β	*SE*	LLCI	ULCI
主效应								
公正世界信念（BJW）	0.24^{**}	0.08	0.07	0.39	0.26^{**}	0.08	0.10	0.41
社会距离（SD）	0.15	0.08	−0.01	0.30	0.14	0.08	−0.02	0.28
调节效应								
BJW×SD					-0.26^{***}	0.06	−0.33	−0.10
Adj-R^2	0.10^{***}				0.16^{***}			
Δ R^2					0.07^{***}			

注：$N = 164$，$^{***}p < 0.001$，$^{**}p < 0.01$。



注：横坐标为社会距离，纵坐标为公正世界信念对规范错觉的回归系数

图 6 社会距离对公正世界信念和规范错觉关系的调节作用

为进一步探讨社会距离在公正世界信念-规范错觉-惩罚行为这一过程中起到的调节作用，使用 PROCESS3.3（Bootstrap $N = 5000$, Model = 7）分析了不同社会距离（$M \pm 1SD$）下公正世界信念对惩罚行为的间接影响（各变量均进行了标准化处理）。结果显示，当社会距离得分低于平均值 1 个标准差（effect = 0.37, $SE = 0.09$, LLCI = 0.19, ULCI = 0.55）或恰好等于平均值（effect = 0.20, $SE = 0.06$, LLCI = 0.07, ULCI = 0.32）时，间接效应的置信区间均不包含 0，但间接效应随着社会距离得分的增加而减小；当社会距离得分高于平均值 1 个标准差（effect = 0.03, $SE = 0.07$, LLCI = −0.12, ULCI = 0.15）时，间接效应的置信区间均包含 0，此时公正世界信念对惩罚行为的间接效应不显著。这一结果说明社会距离能够调节公正世界信念对惩罚行为的间接影响，即存在如图 7 所示的条件过程模型。我们进一步检验了该模型的竞争模型，即同时考虑社会距离对公正世界信念-惩罚行为、规范错觉-惩罚行为的调节效应，使用 PROCESS3.3（Bootstrap $N = 5000$, Model = 59）进行分析后结果显示：将惩罚作为因变量时，公正世界信念×社会距离（$\beta = 0.004$, $t = 0.08$, $p = 0.937$）、规范错觉×社会距离（$\beta = -0.03$, $t = -0.53$, $p = 0.599$）的回归系数不显著，这在很大程度上说明社会距离在上述两个路径中不存在调节效应。

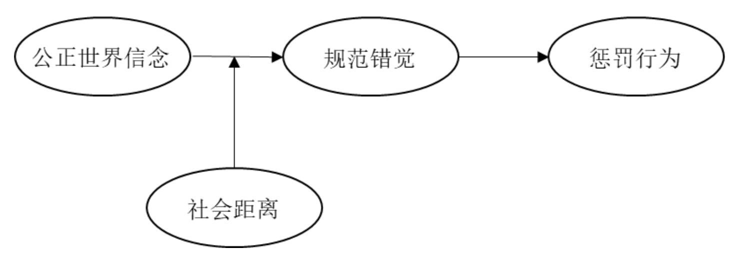


图 7 条件过程模型

综上所述，分析结果较好地支持了假设 3 和假设 4。也就是说，在第三方惩罚中，公正世界信念是带来规范错觉的原因之一，公正世界信念越弱，低估的规范错觉就越强，而自身的惩罚水平就越低，即存在公正世界信念-规范错觉-惩罚行为的中介模型。其次，社会距离能够有效地调节上述模型的前半过程，1）社会距离越远，公正世界信念在上述模型中的作用就越强；2）当社会距离的得分超过 4 时，公正世界信念对规范错觉的影响就不再显著。

5 实验 4

实验 3 从统计分析上表明公正世界信念可以在一定程度上影响被试的规范错觉，但尚无法充分说明两者间的因果关系，因此我们在实验 4 中通过操纵被试的公正世界信念来验证这一因果关系，从而对实验 3 的结果提供有益补充。

5.1 被试

使用软件 G*Power 3.1 进行的功效分析显示：取中等效应量 $f^2 = 0.15$，显著性水平 $\alpha = 0.05$，多元回归至少需要 119 名被试才能达到 95%（$1 - \beta$）的统计检验力；取中等效应量 $f = 0.25$，显著性水平 $\alpha = 0.05$，单因素方差分析至少需要 252 名被试（每组 84 名）才能达到 95%（$1 - \beta$）的统计检验力。按照样本要求量较大的单因素方差分析计算，至少需要 252 名被试，而实际参与实验 4 的被试为 284 名大学生，其中本科生占 81.69%，硕士研究生占 18.31%。被试平均年龄为 21.09 ± 1.78 岁，女性占 51.76%。被试的专业分布如下：理工类占 38.38%，人文类占 22.54%，社科类占 28.52%，艺术及其他占 10.56%。

5.2 变量与设计

实验 4 为 3 组间因子设计，包括激活组（激活公正世界信念）、抑制组（抑制公正世界信念）和对照组。其中公正世界信念（Cronbach's $\alpha = 0.82$）、惩罚（PS）和规范错觉（$PO - M_{PS}$）的测量与实验 3 一致，但实验 4 中被试只需针对 8-2 的分配方案进行估计和实施惩罚，一方面是因为实验 3 中不公平分配方案的不公平程度基本不会影响规范错觉的强弱（$\beta = -0.03, t = -0.70, p = 0.486$），另一方面也是为了和实验 2 的结果相呼应。在计算三组被试的

规范错觉时，减去的平均惩罚（M_{PS}）为组平均值而非总平均值。

5.3 实验程序

实验 4 的范式为带有第三方惩罚的独裁者博弈，通过 z-Tree 上机实验完成（Fischbacher, 2007）。被试被随机分入上述三种实验条件。实验开始后，激活组和抑制组的被试阅读改编自梁福成等（2016）的文本材料，用于激活/抑制公正世界信念，对照组则阅读一段无关材料，三组被试阅读的文本材料均为 147 字符（其中标点符号数量均为 12，详见附录 3）。随后通过前述 6 个题项测试被试的公正世界信念。在充分了解博弈规则（同实验 2）后，被试需估计参加此次实验的其他第三方在面对该分配方案时的平均惩罚，并报告自己的惩罚决策。博弈共 1 轮，实验报酬为出场费加上被试手中剩余的代币数量。

5.4 结果与讨论

不同性别（$t(282) = −1.44$~0.24, $ps = 0.151$~0.813）、教育程度（$t(282) = −0.17$~0.65, $ps = 0.516$~0.946）和专业类型（$F(3, 280) = 0.19$~1.54, $ps = 0.205$~0.905）下被试的公正世界信念、惩罚、估计和规范错觉均无显著差异，且年龄和上述变量之间的相关系数也不显著（$r = −0.05$~$−0.03$, $ps = 0.438$~0.672）。估计和惩罚的配对样本 t 检验结果显示，在抑制组（$t(99) = −6.42$, $p < 0.001$, $d = −0.64$, 95%CI = [−0.55, −0.29]）和对照组（$t(86) = −2.19$, $p = 0.031$, $d = −0.23$, 95%CI = [−0.33, −0.02]）中，存在低估的规范错觉；但在激活组（$t(96) = 0.96$, $p = 0.337$, $BF_{01} = 5.67$）中，规范错觉不显著，即被试对他人惩罚水平的估计和所有被试的实际平均惩罚水平无显著差异，贝叶斯因子分析也说明当前数据更加支持零假设。这意味着在激活公正世界信念后，被试的规范错觉在一定程度上发生了改变。

由于实验 4 的目的在于通过操纵被试的公正世界信念，进而影响规范错觉与惩罚行为，以此说明公正世界信念与规范错觉的因果关系，因此需要通过比较被试间的差异来检验我们对公正世界信念的操纵有效性。对三组被试的公正世界信念、规范错觉和惩罚行为进行了单因素方差分析，结果显示（表 8）：三组被试在上述变量上均存在显著差异。多重比较（Bonferroni 法）的结果表明，激活组的公正世界信念显著高于抑制组（$p < 0.001$, 95%CI = [0.46, 0.95]）和对照组（$p = 0.002$, 95%CI = [0.12, 0.63]），抑制组则显著低于对照组（$p = 0.006$, 95%CI = [−0.59, −0.08]），这说明我们对被试公正世界信念的操纵是有效的。此外，抑制组中存在显著低估的规范错觉，而激活组的规范错觉不显著（尽管单纯从数字上来看，存在高估的规范错觉），且两组间存在显著差异（$p < 0.001$, 95%CI = [0.24, 0.79]）。最后，激活组的惩罚水平也显著高于抑制组（$p = 0.042$, 95%CI = [0.01, 0.62]）。这意味着公正世界信念的激活削弱了规范错觉的低估程度，这为假设 3 提供了额外的证据。图 8 直观地展现了三组被试在

公正世界信念、规范错觉和惩罚行为上的差异。

表 8 公正世界信念、规范错觉和惩罚行为的单因素方差分析

变量	激活组	抑制组	对照组	$F(2, 281)$	p	η^2_p
公正世界信念	4.84 ± 0.63	4.13 ± 0.84	4.46 ± 0.66	23.47	＜0.001	0.143
规范错觉	0.09 ± 0.80	−0.42 ± 0.80	−0.17 ± 0.82	9.92	＜0.001	0.066
惩罚行为	3.13 ± 0.82	2.82 ± 0.94	2.92 ± 0.91	3.18	0.043	0.022

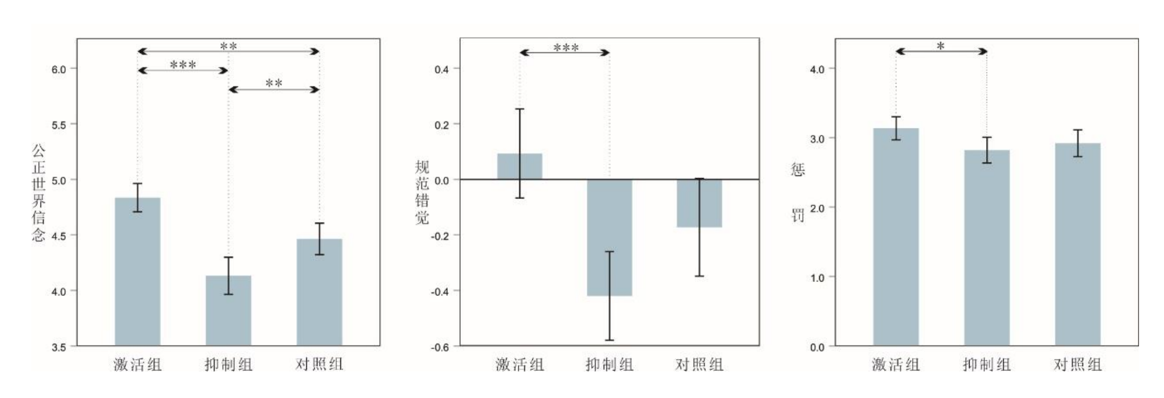


图 8 公正世界信念、规范错觉和惩罚行为的多重比较

注：$^{***}p < 0.001$，$^{**}p < 0.01$，$^{*}p < 0.05$。

上述结果一方面验证了公正世界信念与规范错觉之间的因果关系，另一方面也暗示规范错觉在公正世界信念与惩罚行为之间可能起到中介作用。为了更清楚地说明这一点，使用Preacher 和 Hayes（2004）所开发的 PROCESS3.3 检验公正世界信念-规范错觉-惩罚行为的中介模型（Bootstrap $N = 5000$，Model = 4）。结果显示，公正世界信念对惩罚行为的直接效应不显著（effect = 0.06, SE = 0.06, LLCI = −0.06, ULCI = 0.18），而通过规范错觉影响惩罚行为的间接效应显著（effect = 0.27, SE = 0.05, LLCI = 0.18, ULCI = 0.36）。总体来说，上述分析结果较好地说明了公正世界信念与规范错觉间的因果关系，以及公正世界信念-规范错觉-惩罚行为的中介模型，这为实验 3 的结果提供了更充分的证据。

6 总讨论

6.1 研究结论及其意义

在现有文献中，人际互动对自身惩罚行为的影响尚未受到足够关注。有研究者提出了条

件惩罚这一概念，即个体只有在观察到他人实施惩罚后才会对违规者采取惩罚行为（Huang et al., 2018; Molleman et al., 2019），而 Li 等（2021）将上述影响称为惩罚的社会规范。条件惩罚这一概念预设了一个隐含前提，即个体总是能充分获取他人的惩罚信息。但正如 Kiyonari 和 Barclay（2008）指出，在现实生活中人们往往很难做到这一点。在这种情况下，人们对他人惩罚的估计取代了有关惩罚的真实信息，进而对自身的惩罚行为产生了影响（Kamei, 2020）。Rimal 和 Lapinski（2015）注意到，个体对某个社会规范的感知与真实的规范之间往往存在差异，即所谓的规范错觉（Cox et al., 2019）。本文首次通过实验证明了在第三方惩罚中确实存在规范错觉，并且这种错觉的方向是向下低估的；而需要注意的是，仅在面对不公平的分配方案时，才会出现规范错觉，换言之，人们倾向于低估他人对违规者的惩罚水平。上述结果有两个重要的启发意义。第一，第三方惩罚往往被认为是一种积极行为（Enge et al., 2017），而有关规范错觉的研究指出，人们往往系统性地低估他人的积极行为而高估消极行为（Dempsey et al., 2018），本文的实验结果为上述结论提供了新的佐证。第二，更为重要的是，惩罚与合作在这一点上产生了鲜明的对比：尽管惩罚与合作均被认为是维持人类社会顺利运行的重要力量（Molho et al., 2020; Fehr & Schurtenberger, 2018），但人们却往往低估了他人的惩罚而高估了他人的合作。Fehr 和 Fischbacher（2004）认为策略性动机在对合作的高估中发挥了重要作用，具体而言，因为有利于最大化自身利益，人们策略性地将自身的合作水平控制在平均水平以下，结果表现为人们往往高估了他人的合作水平。从这个角度来说，策略性动机似乎并不能很好地解释本文的实验结果。这一方面表明惩罚主要是为了维护公平（Falk et al., 2005）或是为了缓解由不公平现象而引发的愤怒情绪（Jordan et al., 2016; Pfattheicher et al., 2019），基于利益考虑的策略性动机可能只发挥了次要的作用；另一方面也说明惩罚和合作尽管较为相似，但实际是两种完全不同的社会现象（Molleman et al., 2019），如陈思静和徐烨超（2020）发现，人们倾向于认为合作者既是温暖的（warm）又是有能力的（competent），但对于惩罚者人们只肯定了他们的能力维度，在温暖维度上的评价却是消极的。

其次，我们需进一步挖掘人们低估他人惩罚行为的原因。我们提出了三种解释。第一，由于存在归因偏差，人们总是倾向于低估他人的积极行为（Bursztyn et al., 2020; Dempsey et al., 2018），除非高估他人的积极行为能为自己带来收益。其次，公正世界信念在一定程度上解释了个体对他人惩罚行为的低估。胡金生等（2012）指出，公正判断更多地依赖于道德直觉而非社会线索，因此我们有理由认为，当被试无法获取关于惩罚的外部信息时，规范错觉的产生可能依赖于个体的某种直觉或信念。本文的研究结果显示：较弱的公正世界信念往往

意味着较强的低估的规范错觉。具体来说，被试低估他人惩罚的重要原因可能是，他们对“善有善报，恶有恶报”这一信念的认同度较低，即对违规行为会受到应有惩罚这一观点持怀疑态度，因此人们低估了他人的惩罚水平，而这又进一步减少了被试的惩罚行为。第三，从社会发展的角度来考虑，几乎所有现代社会都处于惩罚权力高度垄断的阶段（Mathew, 2017），即公检法系统或类似的机构垄断了大部分对违规者的惩罚权力，因此尽管“路见不平拔刀相助”式的第三方惩罚行为依然存在，但对于庞大的惩罚系统而言只占到一个较小的比例，这可能在一定程度上加强了人们对他人惩罚行为的低估，因为在现代社会中人们已经习惯了奖惩行为由国家机构出面实施。上述解释提出了一种新的可能性，也就是说比起惩罚权力高度集中的现代社会来，在处于前国家时期的社群中（pre-state community），这种低估是否不存在或程度要小得多？这为未来的研究提供了一种新的理论思路。我们从行为性质、个人信念和社会发展三个角度提出了导致被试低估他人惩罚的原因，但这三个原因是独立的还是相互交织在一起，需要未来更多的研究才能给出可靠的回答。

再次，本文还考察了规范错觉对惩罚的影响。这种影响可以分为两个方面：对感知者（即被试自身）和对被感知群体（即其他惩罚者）的影响，本文重点考察了前者。实验 1 结果表明，对他人惩罚行为的低估导致了被试自身较低的惩罚水平，并且这种影响在控制了不公平程度后仍然成立。为了更好地建立两者的因果关系，我们还在实验 2 中通过有针对性的规范提示改变了被试的规范错觉，结果仍然支持上述观点，即对他人惩罚行为的低估程度越大，被试的惩罚行为就越低。另一方面，低估对被感知群体的影响并不是本文的研究重点，但行为确证理论（behavioral confirmation theory）指出，一旦人们形成了错误的社会信念（在本文中体现为低估了他人的惩罚行为），就有可能引发他人做出某些行为反应以支持这些信念（Snyder & Swann, 1978），换言之，个体对他人惩罚行为的低估可能具有双重影响：一方面降低了自身的惩罚行为，另一面如果抱有这种低估心态的个体足够多，这种低估的信念也可能降低了他人的惩罚水平，从而来迎合这些个体的低估期望。如果上述推理成立，那么对他人惩罚的低估可能最终形成一个双通道的循环系统，从而从总体上降低了第三方惩罚，这也在一定程度上解释了尽管在实验室环境中第三方惩罚屡见不鲜，如在 Fehr 和 Gächter（2002）的经典研究中，74.20%的被试做出了惩罚，但在真实生活中却并不常见（Wu et al., 2016）。从另一方面来说，本文实验结果也为如何克服这种规范错觉提供了可能的途径。我们在最后一个实验中直接操纵了被试的公正世界信念，结果再次说明了公正世界信念与规范错觉间的因果关系，以及公正世界信念-规范错觉-惩罚行为的中介模型。这一结果也具有很强的实践意义：虽然公正世界信念被认为是一种较为稳定的个人特质（吴佩君, 李晔, 2014），但在现

实社会中，我们仍然可以通过恰当的宣传教育提升人们的公正世界信念（Correia et al., 2007; Jiang et al., 2018），进而减小惩罚中低估的规范错觉，增加人们对违规行为的惩罚，以维护社会公平。

最后，本文还揭示了一个重要的现象：感知社会距离调节了公正世界信念对规范错觉的影响，换言之，随着感知社会距离的拉近，被试对他人惩罚行为的估计受公正世界信念的影响也逐步减弱。将感知社会距离视为人与人之间的外部联系（即外在参照点），而将公正世界信念视为一种内部道德直觉（即内在参照点），我们在现有文献基础上（Chen et al., 2021）提出了双重参照点理论：当被试感知到与他人的外部联系较弱时，对其行为的估计主要依赖于自己的道德直觉，在这种情况下个体的公正世界信念预测了个体对他人惩罚行为的估计；而随着个体与他人社会距离的缩小，道德直觉的作用就被削弱了，而一旦超过某个阈值，公正世界信念的预测作用就不再显著。

本研究主要有以下三方面的现实意义。第一，实验 2 表明，通过规范信息提示我们可以在一定程度上影响规范错觉的方向与程度，这意味着在现实中规范信息干预（normative information intervention）可能是影响被试规范错觉进而校正被试行为的一种可行手段。第二，实验 2 和实验 4 的结果为我们提供了两种潜在的干预方式。一方面，虽然在现实中，我们难以针对每一次违规行为进行规范提示，但由于规范提示存在溢出效应（Bursztyn et al., 2020），我们可以通过对现实生活中见义勇为的行为进行及时的宣传报道，进而影响其他违规情境下人们的规范错觉与惩罚行为；另一方面，一种较为简便的干预方式是：通过恰当的宣传教育提升人们的公正世界信念（Correia et al., 2007; Jiang et al., 2018），进而影响规范错觉和惩罚行为。第三，社会距离的调节效应也指出了第二种干预方式的适用范围，即公正世界信念主要是在相对陌生的环境中对规范错觉和惩罚行为发挥作用，因此一个合理的猜想是，在以陌生人为主的公共场合中提升公正世界信念的措施可能会发挥更好的作用。

6.2 研究不足

虽然本文取得了若干有意义的结果，但仍然存在以下不足与局限。第一，本文选用的经济惩罚是第三方惩罚研究中的主流范式，在这种范式下，惩罚成本和违规成本均表现为金钱形式。但仍然存在其他的成本形式的惩罚，例如社会排斥（Liu et al., 2017）和流言（Molho et al., 2020），并且不同的成本形式会对结果产生不同的影响（陈思静等, 2020），未来研究可进一步探究和比较不同惩罚形式下的规范错觉。其次，本文考察的惩罚为第三方惩罚，即惩罚者的利益并未受到违规行为的直接影响，那么在惩罚者的利益与违规行为直接相关的情况下，如最后通牒博弈和公共物品博弈，人们是否仍然会低估他人的惩罚行为？关于这一点本

文无法直接给出答案，未来研究者可以在本文的基础上将研究范式进行拓展而对上述问题给出明确答案。第三，在实验 2 和实验 4 中，我们通过有针对性地改变被试的规范错觉以及操纵其公正世界信念，进而考察了两者对各自结果变量的影响，但需要指出的是，实验结果只能表明在较短的时间内被试的规范错觉和公正世界信念是能够受到外部因素的影响，这种影响是否在较长的时间内依然存在，这需要纵贯研究才能给出答案。第四，在本文的四个实验中，都采用了平均惩罚来代替真实的惩罚规范，进而计算了被试的规范错觉，这也是目前研究中较为普遍的方式（e.g., 陈思静, 濮雪丽等, 2021; Duong & Parker, 2018; Sandstrom et al., 2013）。但平均惩罚和真实的惩罚规范之间可能存在细微的差异，如何更好地衡量存在于集体中的真实规范，是值得未来研究者思考的一个问题。最后，需要说明的是，社会规范对个体行为的影响是一个复杂的认知加工过程。而本文作为探索性研究，主要考察了规范错觉的重要作用与影响因素，并未控制这一过程中的诸多影响因素，如何识别这些因素并加以控制是未来研究值得探讨的一个重要方向。

参考文献

Amialchuk, A., Ajilore, O., & Egan, K. (2019). The influence of misperceptions about social norms on substance use among school-aged adolescents. *Health Economics*, *28*(6), 736–747.

Anthenien, A. M., DeLozier, S. J., Neighbors, C., & Rhodes, M. G. (2018). College student normative misperceptions of peer study habit use. *Social Psychology of Education*, *21*(2), 303–322.

Bai, B. Y., Liu, X. X., & Kou, Y. (2014). Belief in a just world lowers perceived intention of corruption: The mediating role of perceived punishment. *PloS one*, *9*(5), e97075.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, *7*, 13327

Bègue, L., & Bastounis, M. (2003). Two spheres of belief in justice: Extensive support for the bidimensional model of belief in a just world. *Journal of Personality*, *71*(3), 435–463.

Bicchieri, C., Dimant, E., & Gächter, S. (2020). *Observability, social proximity, and the erosion of norm compliance*. (CESifo Working Paper No. 8212). Munich, Germany: Center for Economic Studies and ifo Institute.

Bouman, T., & Steg, L. (2019). Motivating society-wide pro-environmental change. *One Earth*, *1*(1), 27–30.

Bouman, T., Steg, L., & Zawadzki, S. J. (2020). The value of what others value: When perceived biospheric group values influence individuals' pro-environmental engagement. *Journal of Environmental Psychology*, *71*, 101470.

Bursztyn, L., González, A. L., & Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *The American Economic Review*, *110*(10), 2997–3029.

Burton-Chellew, M. N., El Mouden, C., & West, S. A. (2017). Social learning and the demise of costly cooperation in humans. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1853), 20170067.

Charness, G., & Gneezy, U. (2008). What's in a name? Anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior & Organization*, *68*(1), 29–35.

Chen, H., Zeng, Z., & Ma, J. (2020). The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions. *PloS One*, *15*(3), e0229510.

Chen, S. J., Hu, H. M., & Yang, S. S. (2020). Payment vs. Retaliation: Impact of Cost Form on Third-Party Punishment. *Journal of Psychological Science*, *43*(02), 416–422.

[陈思静, 胡华敏, 杨莎莎. (2020). 支付与报复: 成本形式对第三方惩罚的影响. *心理科学, 43*(02), 416–422.]

Chen, S. J., Pu, X. L., Zhu, Y., Wang, H., & Liu, J. W. (2021). The impact of normative misperception on food waste in dining out: Mechanism analyses and countermeasures. *Acta Psychologica Sinica*, *53*(8), 904–918.

[陈思静, 濮雪丽, 朱玥, 汪昊, 刘建伟. (2021). 规范错觉对外出就餐中食物浪费的影响：心理机制与应对策略. *心理学报. 53*(8), 904–918.]

Chen, S. J., Xing, Y. L., Weng, Y. J., & Li, C. (2021). Spillover effects of third-party punishment on cooperation: A norm-based explanation. *Acta Psychologica Sinica*, *53*(7), 758–772.

[陈思静, 邢懿琳, 翁异静, 黎常. (2021). 第三方惩罚对合作的溢出效应: 基于社会规范的解释.*心理学报*, *53*(7), 758–772.]

Chen, S. J., & Xu, Y. C. (2020). Warmth and competence: Impact of third-party punishment on punishers' reputation. *Acta Psychologica Sinica*, *52*(12), 1436–1451.

[陈思静, 徐烨超. (2020). "仁者"还是"智者": 第三方惩罚对惩罚者声誉的影响. *心理学报*, *52*(12), 1436–1451.]

Chen, S., Liu, J., & Hu, H. (2021). A norm-based conditional process model of the negative impact of optimistic bias on self-protection behaviors during the COVID-19 pandemic in three Chinese cities. *Frontiers in Psychology*, *12*, 659218.

Correia, I., Vala, J., & Aguiar, P. (2007). Victim's innocence, social categorization, and the threat to the belief in a just world. *Journal of Experimental Social Psychology*, *43*(1), 31–38.

Cox, M. J., DiBello, A. M., Meisel, M. K., Ott, M. Q., Kenney, S. R., Clark, M. A., & Barnett, N. P. (2019). Do misperceptions of peer drinking influence personal drinking behavior? Results from a complete social network of first-year college students. *Psychology of Addictive Behaviors*, *33*(3), 297–303.

Csukly, G., Polgár, P., Tombor, L., Réthelyi, J., & Kéri, S. (2011). Are patients with schizophrenia rational maximizers? Evidence from an ultimatum game study. *Psychiatry Research*, *187*(1-2), 11–17.

Davis, J. P., Pedersen, E. R., Tucker, J. S., Dunbar, M. S., Seelam, R., Shih, R., & D'Amico, E. J. (2019). Long-term associations between substance use-related media exposure, descriptive norms, and alcohol use from adolescence to young adulthood. *Journal of Youth and Adolescence*, *48*(7), 1311–1326.

de Kwaadsteniet, E. W., van Dijk, E., Wit, A., De Cremer, D., & de Rooij, M. (2007). Justifying decisions in social dilemmas: Justification pressures and tacit coordination under environmental uncertainty. *Personality and Social Psychology Bulletin*, *33*(12), 1648–1660.

Dempsey, R. C., McAlaney, J., & Bewick, B. M. (2018). A critical appraisal of the social norms approach as an interventional strategy for health-related behavior and attitude change. *Frontiers in psychology*, *9*, 2180.

Dillon, C. E., & Lochman, J. E. (2019). Correcting for norm misperception of anti-bullying attitudes. *International Journal of Behavioral Development*. DOI: https://doi.org/10.1177/0165025419860598.

Dumas, T. M., Davis, J. P., & Neighbors, C. (2019). How much does your peer group really drink? Examining the relative impact of overestimation, actual group drinking and perceived campus norms on university students' heavy alcohol use. *Addictive Behaviors*, *90*, 409–414.

Duong, H. T., & Parker, L. (2018). Going with the flow. *Journal of Social Marketing*, *8*(3), 314–332.

Enge, S., Mothes, H., Fleischhauer, M., Reif, A., & Strobel, A. (2017). Genetic variation of dopamine and serotonin function modulates the feedback-related negativity during altruistic punishment. *Scientific Reports*, *7*, 2996.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*(6), 2017–2030.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458–468.

FeldmanHall, O., Otto, A. R., & Phelps, E. A. (2018). Learning moral values: Another's desire to punish enhances one's own punitive behavior. *Journal of Experimental Psychology: General*, *147*(8), 1211–1224.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, *100*(1), 541–556.

Gächter, S., Kölle, F., & Quercia, S. (2017). Reciprocity and the tragedies of maintaining and providing the commons. *Nature Human Behaviour*, *1*(9), 650–656.

Ganz, G., Neville, F. G., Kassanjee, R., & Ward, C. L. (2020). Parental misperceptions of in-group norms for child discipline. *Journal of Community & Applied Social Psychology*, *30*(6), 628–644.

Goeschl, T., Kettner, S. E., Lohse, J., & Schwieren, C. (2018). From social information to social norms: Evidence from two experiments on donation behaviour. *Games*, *9*(4), 91–115.

Grimm, V., Utikal, V., & Valmasoni, L. (2017). In-group favoritism and discrimination among multiple out-groups.

Journal of Economic Behavior & Organization, *143*, 254–271.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*(1), 79–89.

Hu, C. P., Kong, X. Z., Wagenmakers, E. J., Ly, A., & Peng, K. P. (2018). The Bayes factor and its implementation in JASP: A practical primer. *Advances in Psychological Science*, *26*(6), 951–965.

[胡传鹏, 孔祥祯, Wagenmakers, E. J., Ly, A., 彭凯平. (2018). 贝叶斯因子及其在 JASP 中的实现. *心理科学进展*, *26*(6), 951–965.]

Hu, J. S., Ye, C., Li, X., & Gao, T. T. (2012). "Irrationality" in justice judgment: Processing mechanisms, main forms and influencing factors. *Advances in Psychological Science*, *20*(05), 726–734.

[胡金生, 叶春, 李旭, 高婷婷. (2012). 公正判断中的"非理性": 加工特征、主要表现和影响因素. *心理科学进展*, *20*(05), 726–734.]

Huang, F., Chen, X., & Wang, L. (2018). Conditional punishment is a double-edged sword in promoting cooperation. *Scientific Reports*, *8*, 528.

Ji, W. H., Zhang, L. G., & Kou, Y. (2014). How the Belief in a Just World Influence College Student's Intention to Help People in Need: The Role of Attribution of Responsibility and the Cost of Helping. *Psychological Development and Education*, *30*(05),496–503.

[姬旺华, 张兰鸽, 寇彧. (2014). 公正世界信念对大学生助人意愿的影响:责任归因和帮助代价的作用. *心理发展与教育*, *30*(05),496–503.]

Jiang, R., Liu, R. D., Ding, Y., Zhen, R., Sun, Y., & Fu, X. (2018). Teacher justice and students' class identification: Belief in a just world and teacher-student relationship as mediators. *Frontiers in psychology*, *9*, 802.

Jones, P. E. (2004). False consensus in social context: Differential projection and perceived social distance. *British Journal of Social Psychology*, *43*(3), 417–429.

Jordan, J. J., McAuliffe, K., & Rand, D. G. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, *19*(4), 741–763.

Kamei, K. (2020). Group size effect and over-punishment in the case of third-party enforcement of social norms. *Journal of Economic Behavior & Organization*, *175*, 395–412.

Keizer, K., & Schultz, P. W. (2018). Social norms and pro-environmental behaviour. In L. Steg, & J. I. M. de Groot (Eds.), *Environmental psychology: An introduction* (pp. 179–188). Chichester, UK: John Wiley & Sons.

Kenney, S. R., Ott, M., Meisel, M. K., & Barnett, N. P. (2017). Alcohol perceptions and behavior in a residential peer social network. *Addictive Behaviors*, *64*, 143–147.

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, *95*(4), 826–842.

Kosfeld, M., & Rustagi, D. (2015). Leader punishment and cooperation in groups: Experimental field evidence from commons management in Ethiopia. *American Economic Review*, *105*(2), 747–783.

Laninga-Wijnen, L., Harakeh, Z., Dijkstra, J. K., Veenstra, R., & Vollebergh, W. (2018). Aggressive and prosocial peer norms: Change, stability, and associations with adolescent aggressive and prosocial behavior development. *The Journal of Early Adolescence*, *38*(2), 178–203.

Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology*, *58*(1), 317–344.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D., & Sutter, M. (2014). Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(19), 6916–6921.

Lerner, M. J. (1965). Evaluation of performance as a function of performer's reward and attractiveness. *Journal of Personality and Social Psychology*, *1*(4), 355–360.

Lerner, M. J., & Miller, D. T. (1978). Just world research and the attribution process: Looking back and ahead.

Psychological Bulletin, *85*(5), 1030–1051.
Liang, F. C., Zhou, Y., Wang, J. K., & Tang, W. H. (2016). Just world beliefs and motivation effect of cross-context. *Studies of Psychology and Behavior*, *14*(03), 367–371+383.
[梁福成, 周宇, 王俊坤, 唐卫海. (2016). 公正世界信念与跨情境动机效应. *心理与行为研究*, *14*(03), 367–371+383.]
Li, J., Li, S., Wang, P., Liu, X., Zhu, C., Niu, X., ... & Yin, X. (2018). Fourth-Party Evaluation of Third-Party Pro-social Help and Punishment: An ERP Study. *Frontiers in Psychology*, *9*, 932.
Li, M. H., & Rao, L. L. (2017). Moral judgment from construal level theory perspective. *Advances in Psychological Science*, *25*(08), 1423–1430.
[李明晖, 饶俪琳. (2017). 解释水平视角下的道德判断. *心理科学进展*, *25*(08), 1423–1430.]
Liu, G. Z., Zhang, D. J., Zhu, Z. G., Li, J. J., & Chen, X. (2020). The effect of family socioeconomic status on adolescents' problem behaviors: The chain mediating role of parental emotional warmth and belief in a just world. *Psychological Development and Education*, *36*(02), 240–248.
[刘广增, 张大均, 朱政光, 李佳佳, 陈旭. (2020). 家庭社会经济地位对青少年问题行为的影响:父母情感温暖和公正世界信念的链式中介作用. *心理发展与教育*, *36*(02), 240–248.]
Liu, L., Chen, X., & Szolnoki, A. (2017). Competitions between prosocial exclusions and punishments in finite populations. *Scientific Reports*, *7*, 46634.
Li, X., Molleman, L., & van Dolder, D. (2021). Do descriptive social norms drive peer punishment? Conditional punishment strategies and their impact on cooperation. *Evolution and Human Behavior*, *42*(5), 469–479.
Mathew, S. (2017). How the second-order free rider problem is solved in a small-scale society. *American Economic Review*, *107*(5), 578–581.
Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(28), 11375–11380.
Molho, C., Tybur, J. M., Van Lange, P. A., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications*, *11*, 3432.
Molleman, L., Kölle, F., Starmer, C., & Gächter, S. (2019). People prefer coordinated punishment in cooperative interactions. *Nature Human Behaviour*, *3*(11), 1145–1153.
Neugebauer, T., Perote, J., Schmidt, U., & Loos, M. (2009). Selfish-biased conditional cooperation: On the decline of contributions in repeated public goods experiments. *Journal of Economic Psychology*, *30*(1), 52–60.
Ozono, H., Jin, N., Watabe, M., & Shimizu, K. (2016). Solving the second-order free rider problem in a public goods game: An experiment using a leader support system. *Scientific Reports*, *6*, 38349.
Pfattheicher, S., Sassenrath, C., & Keller, J. (2019). Compassion magnifies third-party punishment. *Journal of Personality and Social Psychology*, *117*(1), 124–141.
Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717–731.
Przepiorka, W., & Liebe, U. (2016). Generosity is a sign of trustworthiness—the punishment of selfishness is not. *Evolution and Human Behavior*, *37*(4), 255-262.
Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, *27*(9), 1192–1206.
Rau, R., Nestler, S., Geukes, K., Back, M. D., & Dufner, M. (2019). Can other-derogation be beneficial? Seeing others as low in agency can lead to an agentic reputation in newly formed face-to-face groups. *Journal of Personality and Social Psychology*, *117*(1), 201–227.
Rimal, R. N., & Lapinski, M. K. (2015). A re-explication of social norms, ten years later. *Communication Theory*, *25*(4), 393–409.

Sandstrom, M., Makover, H., & Bartini, M. (2013). Social context of bullying: Do misperceptions of group norms influence children's responses to witnessed episodes? *Social Influence*, *8*(2-3), 196–215.

Sasaki, T., Uchida, S., & Chen, X. (2015). Voluntary rewards mediate the evolution of pool punishment for maintaining public goods in large populations. *Scientific Reports*, *5*, 8917.

Sawitri, D. R., Hadiyanto, H., & Hadi, S. P. (2015). Pro-environmental behavior from a socialcognitive theory perspective. *Procedia Environmental Sciences*, *23*, 27–33.

Schlag, K. H., Tremewan, J., & Van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, *18*(3), 457–490.

Snyder, M., & Swann, W. B. (1978). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology*, *14*(2), 148–162.

Son, J. Y., Bhandari, A., & FeldmanHall, O. (2019). Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Scientific Reports*, *9*, 11625.

Strelan, P., Di Fiore, C., & Prooijen, J. W. V. (2017). The empowering effect of punishment on forgiveness. *European Journal of Social Psychology*, *47*(4), 472–487.

Sutton, R. M., & Winnard, E. J. (2007). Looking ahead through lenses of justice: The relevance of just-world beliefs to intentions and confidence in the future. *British Journal of Social Psychology*, *46*(3), 649–666.

Testé, B., & Perrin, S. (2013). The impact of endorsing the belief in a just world on social judgments. *Social Psychology*, *44*(3), 209–218.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463.

Tumasjan, A., Strobel, M., & Welpe, I. (2011). Ethical leadership evaluations after moral transgression: Social distance makes the difference. *Journal of Business Ethics*, *99*(4), 609–622.

Volk, S., Nguyen, H., & Thöni, C. (2019). Punishment under threat: The role of personality in costly punishment. *Journal of Research in Personality*, *81*, 47–55.

Wallace, B., Cesarini, D., Lichtenstein, P., & Johannesson, M. (2007). Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(40), 15631–15634.

Wen, Z. L., & Ye, B. J. (2014). Analyses of mediating effects: The development of methods and models. *Advances in Psychological Science, 22*(5), 731–745.

[温忠麟, 叶宝娟. (2014). 中介效应分析: 方法和模型发展. *心理科学进展*, *22*(5), 731–745.]

Wu, J., Balliet, D., & van Lange, P. A. (2016). Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific Reports*, *6*, 23919.

Wu, J. S., & Wang, N. (2017). The effect of social distance on sexual attribution bias of success or failure. *Journal of Psychological Science*, *40*(05),1222–1227.

[吴静珊, 王娜. (2017). 社会距离对成败行为性别归因偏差的影响. *心理科学*, *40*(05),1222–1227.]

Wu, M. S., Yan, X., Zhou, C., Chen, Y., Li, J., Zhu, Z., … Han, B. (2011). General belief in a just world and resilience: Evidence from a collectivistic culture. *European Journal of Personality*, *25*(6), 431–442.

Wu, P. J., & Li, Ye. (2014). Cultural differences of the belief in a just world. *Advances in Psychological Science*, *22*(11), 1814–1822.

[吴佩君, 李晔. (2014). 公正世界信念的文化差异. *心理科学进展*, *22*(11), 1814–1822.]

Xu, J., Sun, X. C., Dong, Y., Wang, Z. J., Li, W. Q., & Yuan, B. (2017). Compensation or Punishment——the Effect of Social Distance on Third-party Intervention. *Journal of Psychological Science*, *40*(05),1175–1181.

[徐杰, 孙向超, 董悦, 汪祚军, 李伟强, 袁博. (2017). 人情与公正的抉择:社会距离对第三方干预的影响. *心理科学*, *40*(05),1175–1181.]

Zhang, Y., Pan, Z., Li, K., & Guo, Y. (2018). Self-serving bias in memories. *Experimental Psychology*, *65*(4), 236–244.

Zhou, C. Y., & Guo, Y. Y. (2013). Belief in a just world: A double-edged sword for justice restoration. *Advances in Psychological Science*, *21*(1), 144–154.

[周春燕, 郭永玉. (2013). 公正世界信念——重建公正的双刃剑. *心理科学进展*, *21*(1),144–154.]

Normative misperception in third-party punishment: An explanation from the perspective of belief in a just world

YANG Shasha[1], CHEN Sijing[2]

([1] School of Economics, Shanghai University, Shanghai, 200444, China)

([2] School of Economics and Management, Zhejiang University of Science and Technology, Hangzhou 310023, China)

Abstract

Punishment decisions might be guided by the norm of punishment, that is, people will implement their own punishment according to perceived prevalence of punishment in a similar social midst. However, there may be differences between an individual's perception of norms and actual norms, which is called normative misperception. This article uses four experiments to explore the existence, the direction, and the cause of the normative misperception in third-party punishment, as well as its influence on people's own punitive behaviors.

In Experiment 1, 449 participants were randomized in a four group factorial design (punishing before estimating, estimating before punishing, punishing only, and estimating only). Experiment 1 consisted of 6 rounds of dictator game, in which participants made punishment decisions for 6 offers and/or estimated the average punishment level of other participants in each offer. Experiment 2 aimed to establish the causal relationship between the normative misperception and the punishment by directly manipulating the normative misperception. Specifically, 134 participants were randomly divided into the overestimation group and underestimation group. After receiving the feedback, participants made punishment decision for an unfair offer and estimated the level of punishment of others in this offer. The purpose of Experiment 3 was to test the model of belief in a just world (BJW)-normative misperception-punishment, as well as the moderating effect of perceived social distance (PSD), with a within-participants design involving 164 participants. The procedure was similar to that of Experiment 1, except that we measured participants' BJW and PSD before and after the game, respectively. In Experiment 4, we manipulated participants' BJW through reading materials to test the causal relationship between BJW and the normative misperception.

The results of Experiment 1 showed that there is an underestimated normative misperception in third-party punishment, which leads to a lower level of punishment. Experiment 2 proved that there exists a causal relationship between the normative misperception and punishment by directly

manipulating the independent variables. Experiment 3 demonstrated that BJW might be an underlying cause of the normative misperception, while PSD moderates the effect of BJW on the normative misperception. Finally, Experiment 4 showed the causal relationship between BJW and the normative misperception, providing additional evidence to the results of Experiment 3.

To sum up, we have found evidence of normative misperception in third-party punishment through 4 experiments. This underestimated misperception might be affected by dual reference points: BJW (internal) and PSD (external). It also shows to a certain extent that third-party punishment is a norm-maintaining behavior rather than a gain-based strategic behavior.

Key words third-party punishment, normative misperception, belief in a just world, perceived social distance

附录 1

1. 我认为这个世界基本上是公正的。

2. 在很大程度上，我相信人们得到了他们应得的。

3. 我确信公正总是可以战胜不公正。

4. 从长远来说，我相信遭受不公正的人将会得到补偿。

5. 我坚信在生活的各个领域（包括职业、家庭、政治等方面）里，不公正是偶然的，而不是必然。

6. 我认为人们在做重大决定时会力求公正。

附录 2

1. 总的来说，我认为我可以和他们相处得很好。
2. 我认为我和他们存在一些共同之处。
3. 在有些事情上，我会和他们有相同的看法。
4. 如果我在生活中遇到他们，我会和他们愉快地相处。

附录 3

1. 激活组

一位老人被卡车撞伤，躺在地上不省人事，肇事司机擅自逃离事故现场，在老人生命垂危之际，好心的陈先生将老人及时送往医院进行救治。由于陈先生的帮助，老人得到了及时的救治，老人的家人也对陈先生表示感谢。在监控录像以及老人回忆的帮助下，公安机关抓获逃逸司机，该司机被判有期徒刑 1 年，并对老人一家予以赔偿。

2. 抑制组

一位老人被卡车撞伤，躺在地上不省人事，肇事司机擅自逃离事故现场，在老人生命垂危之际，好心的陈先生将老人及时送往医院进行救治。然而，老人的家人认为陈先生就是肇事司机，并要求赔偿，老人也在家人的压力下违心指认陈先生就是肇事者。由于缺乏目击证人，且现场处于监控盲区，法官判处陈先生对老人予以相应赔偿。

3. 对照组

世界各大洋的主要洋流分布与风带有着密切的关系，但洋流流动的方向和风向一致，在北半球向右偏，南半球向左偏。在热带、亚热带地区，北半球的洋流基本上是围绕副热带高气压作顺时针方向流动，在南半球作逆时针方向流动。在热带由于信风把表层海水向西吹，形成了赤道洋流。东西方向流动的洋流遇到大陆，便向南北分流。